Transforming the AnCora corpus to HPSG

Luis Chiruzzo, Dina Wonsever
{luischir,wonsever}@fing.edu.uy
Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay

February 2016

Abstract

We present the construction of a HPSG corpus for Spanish, based on the transformation of the AnCora Spanish corpus into a HPSG compatible format. We describe the transformation process and the evaluation of the resulting corpus.

1 Introduction

We describe a partial result of a currently ongoing project for building a statistical HPSG parser for Spanish. In this work we transform the AnCora Spanish corpus from its CFG-style annotations to a HPSG compatible format. Head-driven Phrase Structure Grammars (HPSG) [1] are a strongly lexicalized grammar formalism. This family of grammars are very expressive, allowing the modeling of many linguistic phenomena and capturing syntactic and semantic notions at the same time. The rules used in a HPSG grammar are very generic, inspired by X' theory, indicating how a syntactic head can be combined with its complements, modifiers (adjuncts) and specifier. The categories of the elements are organized in a type hierarchy and the parsing result is a tree whose nodes are typed feature structures [2].

Our work is inspired on Enju [3], a statistical HSPG parser for English that has high performance and language coverage. This parser was built based on the Penn Treebank corpus [4]. As the Penn Treebank is not annotated in a HPSG compatible format but rather a CFG-style grammar, they built a set of rules to transform the Penn Treebank trees into a structure that is similar to HPSG [5]. The Enju parser is trained using the result of this transformation.

Other HPSG grammars for Spanish exist, the most relevant one being the Spanish Resource Grammar (SRG) [6], a Spanish HPSG grammar built using the LinGO Grammar Matrix [7], a framework for building HPSG grammars for many languages. SRG uses rule based parser LKB [8], and the results are very rich HPSG trees that include all of the constructions supported by the theory.

Our objective is to build a new HPSG parser whose representations will not be as rich as SRG's, but we aim at making it faster and more robust.

There is no Spanish corpus with HPSG annotations available, but there are corpora with syntactic annotations that could be transformed into something compatible with HPSG. AnCora is a corpus for Spanish and Catalan [9] that contains about half a million words in 17,000 sentences. The corpus has CFG-style annotations, but it is also enriched with attributes such as morphological information and arguments structure.

2 Development

In a HPSG tree, it is necessary to know the syntactic head of every constituent and also the roles that the rest of the elements of the constituents have. This information is not directly available in AnCora, so we created a series of heuristics that exploit the information in the corpus (structure and attributes) in order to transform it to a HPSG compatible format. If we consider the syntactic structures of AnCora as annotated in a CFG, the number of rules in this grammar would be very large. For example, there are 5,800 ways of writing a subordinate sentence, and 900 ways of writing noun phrases. Because of this, we tried to reduce the complexity of the problem using a transformation process which uses two stages: a top-down approach that works together with a bottom-up approach.

We define an *elementary HPSG tree* as a simple tree that consist of a syntactic head surrounded by elements that are directly related to the head (complements, modifiers, specifier). The top-down process tries to transform the most complex structures of the corpus into simpler trees. This means breaking up a node with too many children into a composition of elementary trees that preserve the original structure. The top-down process is in charge, among other things, of extracting quoted or punctuated blocks; marking clitics; extracting prepositional phrases, relative clauses and subordinate sentences; and binarizing sequences of coordinations.

The bottom-up approach assumes that the top-down process has dealt with all those complex structures and left only a set of homogeneous simpler structures, those structures will become elementary HPSG trees after the transformation. In order to transform these trees, we created head detection and arguments classification heuristics. For the English language there is a commonly used heuristic for finding the syntactic head of a phrase in the Penn Treebank corpus, as described in [10]. As there is no equivalent for Spanish, and the grammatical differences between both languages make it impossible to apply the same rules, a set of head detection rules was manually crafted for the elements of Ancora. We defined lists of constraints that an element must match in order to be considered the head of a constituent. The constraints are written in a small language for rules that was created for this purpose. Table 1 shows some examples of the list of detection rules that is used to find the head of a noun phrase (elements of type `grup.nom` in AnCora).

- `n` (noun)
  "...Río_Bravo y Saltillo para la [[$_{\text{H}}$ compañía] [francesa]]..."
- `grup.nom` (nested noun phrase)
  "...y sobre [[$_{\text{H}}$ transmisiones y retenciones] [de fondos de inversión]] ."
- `p` (pronoun)
  "...obtuvo 19 diputados, [[$_{\text{H}}$ dos] [más]] que en 1996..."
- `w` (date)
  "...hundimiento del "Kursk" el [[pasado] [$_{\text{H}}$ 12_de_agosto]] en aguas árticas..."
- `z` (number)
  "...donde lograron el [[$_{\text{H}}$ 71_por_ciento] [de los sufragios]] ..."
- `a` (adjective)
  "...quien cuestiona al entrenador es [[$_{\text{H}}$ enemigo] [del Barça]] ."
- `v` (verb)
  "...sobre todo en el [[$_{\text{H}}$ capitulo] [de las infraestructuras]] ..."
- `s.a` (adjective phrase)
  "...y la [[$_{\text{H}}$ segunda] [, mucho más potente,]] a las 07.30.42..."
- `participi` (participle)
  "...el relato ZZadjNM de lo [[$_{\text{H}}$ ocurrido] [en la sima de ZZlugar]] ..."
- `S/clausetype=participle` (subordinate sentence of type participle)
  "...en_lugar_del [[$_{\text{H}}$ destituido] [Carlos_Sainz_de_Aja]] ."
- `S/clausetype=relative` (subordinate sentence of type relative)
  "...incluidos los [[$_{\text{H}}$ que él mismo ha hablado] [sobre sí mismo]] ..."
- `S/clausetype=completive` (subordinate sentence of type completive)
  "Al [[$_{\text{H}}$ correr] [de los siglos]] se había manifestado un..."
- `sp` (prepositional phrase)
  "aeropuerto de Miami, uno de los [[$_{\text{H}}$ de mayor tráfico aéreo] [en EEUU]]..."
- `sn` (noun phrase, maximal projection)
  "...el hotel (un [[$_{\text{H}}$ cinco estrellas de gran lujo]])..."

Table 1: Rules for head detection inside a `grup.nom`

After finding the syntactic head of a phrase, we proceed to analyze the elements that are directly to the left or to the right of the head, and apply a series of heuristics that try to classify the role of those arguments with respect to the head. The heurisitcs use information about the node such as its part of speech, but also the attributes of the element. The rules for classifying the elements are written in the same language as the rules for detecting heads. In total there are 70 head detection rules and 184 argument classification rules.

Besides these rules, there are specific transformation heuristics for verb phrases, because the verb phrases in AnCora behave in a different way than other constituents and could not be reduced to elementary HPSG trees.

3 Evaluation

The transformed corpus contains only binary or unary constituents and all nodes indicate their syntactic head and the applied rule. AnCora has a total of 780950 constituents and almost all of them could be transformed. We evaluated the accuracy of the transformation heuristics in the following way: We took a random sample of 40 sentences (779 constituents) and manually identified the syntactic head of every constituent and the role of every other element with respect to the head (complement, modifier, specifier, clitic or punctuation mark).

We found that the head detection heuristics have a precision of 95.3%, which climbs to 98.7% if we do not consider the nodes with coordinations. Table 2 shows the precision of the head detection rules by constituent category, considering nodes with coordinations.

Category	Total	Correct	Precision
grup.a	9	6	66.7%
grup.adv	3	3	100.0%
grup.nom	162	154	95.1%
grup.verb	23	23	100.0%
infinitiu	3	3	100.0%
relatiu	1	1	100.0%
S	91	85	93.4%
s.a	4	3	75.0%
sa	1	1	100.0%
sadv	7	7	100.0%
sentence	40	35	87.5%
sn	220	216	98.2%
sp	207	204	98.6%
spec	8	1	12.5%

Table 2: Precision of head detection rules

The arguments classification heuristics have a precision of 92.5% on average, and the category which is the most difficult to classify is the complements (84.95% precision). Table 3 shows the confusion matrix for the arguments classification.

	Specifier	Complement	Modifier	Clitic	Punctuation
Specifier	279	3	3	0	0
Complement	6	333	53	0	0
Modifier	1	18	247	0	0
Clitic	0	0	0	19	0
Punctuation	0	0	0	0	155

Table 3: Confusion matrix for the arguments classification

References

[1] Carl Pollard and Ivan A Sag. *Head-driven phrase structure grammar.* University of Chicago Press, 1994.

[2] Bob Carpenter. *The logic of typed feature structures.* Cambridge University Press, 1992.

[3] Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Efficient hpsg parsing with supertagging and cfg-filtering. In *IJCAI*, pages 1671–1676, 2007.

[4] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[5] Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. In *Natural Language Processing–IJCNLP 2004*, pages 684–693. Springer, 2005.

[6] Montserrat Marimon. The spanish resource grammar. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*, pages 17–23, Valletta, Malta, 2010.

[7] Emily M Bender, Dan Flickinger, and Stephan Oepen. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15*, pages 1–7. Association for Computational Linguistics, 2002.

[8] Ann Copestake, John Carroll, Rob Malouf, and Stephan Oepen. The (new) lkb system. *Center for the Study of Language and Information, Stanford University*, 1999.

[9] Mariona Taulé, Maria Antònia Martí, and Marta Recasens. Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC*, 2008.

[10] Michael Collins. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637, 2003.